RISK: Health, Safety & Environment (1990-2002)

Volume 1
Number 2 *RISK: Issues in Health & Safety*

Article 7

March 1990

Scientific Conventions, Ethics and Legal Institutions

Carl F. Cranor

Scientific Conventions, Ethics and Legal Institutions

Carl F. Cranor*

Introduction

In the early 1970's Congress passed a number of environmental health laws that required federal executive branch agencies to take a prospective approach in trying to prevent health hazards from arising as a consequence of exposure to toxic substances.[1] Many, but not all these statutes permit (and some may require) that the agencies establish risks likely to be presented by exposure to toxic substances in the environment, and, then in accordance with the authority of the statute in question, to regulate those risks.[2] Largely at the instigation of William Ruckleshaus of the Environmental Protection Agency, regulatory agencies began making a distinction for regulatory purposes between risk assessment and risk management. Risk assessment is the "characterization of potential adverse health effects of human exposures to environmental hazards...".[3]

---

* Professor Cranor chairs the Department of Philosophy at the University of California, Riverside. He has his Ph.D. from the University of California, Los Angeles and an M.S.L. from Yale Law School.

1 *See e.g.*, section 306 (d) of The Clean Water Act and section 504 of The Clean Water Act. Federal Water Pollution Control Act Amend. of 1972, Pub. L. No. 92-500, 86 Stat. 816 (1972) (codified at 33 U. S. C. § 1251 (1972)).

2 Some statutes, such as the Delaney Clause [Food Additives Amend. of 1958, Pub. L. No. 85-929, § 409 (c) (3) (A), 72 Stat. 1784 (1958) (codified at 21 U . S. C. § 348 (c) (3) (A))] of The Federal Food, Drug & Cosmetic Act, however, may substantially limit agency discretion.

3 Risk assessments include several elements: description of the potential adverse health effects based on an evaluation of the results of

Typically risk assessment is contrasted with risk management which the National Academy of Sciences indicates is the "process of evaluating alternative regulatory actions and selecting among them."[4] This selection necessarily requires the use of value judgments on such issues as the acceptability of risk and the reasonableness of the cost of control.[5]

Although both the National Academy of Sciences and Administrator Ruckleshaus draw a sharp distinction between risk assessment and risk management, one thesis of this paper is that this is an artificial, and, I believe, an untenable distinction in the present circumstances of scientific uncertainty of carcinogen risk assessments. In addition, in many respects the *scientific* aspects of risk assessment *cannot* be separated from essential social policy judgments that are needed to design and use the scientific data for regulatory and other legal purposes. Thus, risk assessment is necessarily infected with risk management kinds of considerations. This in turn should affect our approach to risk assessment and the law.

---

epidemiological, clinical, toxicologic and environmental research; extrapolation from those results to predict the type and estimate the extent of health effects in humans under given conditions of exposure; judgement as to the number of characteristics of persons exposed at various intensities in duration; and some are judged on the existence of overall magnitude of the public health problem. Risk assessment also includes characterization of uncertainties inherent in the process of inferring risks.

THE NATIONAL RESEARCH COUNCIL, RISK ASSESSMENT IN THE FEDERAL GOVERNMENT: MANAGING THE PROCESS 18 (1983).

4 This is an agency decision making process that entails consideration of political, social economic and engineering information with risk related information to develop, analyze and compare regulatory options and to select the appropriate regulatory response to a potential chronic health hazard.

Id., at 18 and 19.

5 *Id.*

I

Here I will discuss some of the reasons that the distinctions between the "science" of risk assessment and risk management are untenable. I focus on cancer risk assessment for this is probably the most studied and perhaps one of the most controversial areas of risk assessment; thus most is known about it.

A. Because of the large numbers of uncertainties in present risk assessment practices, assessors can dominate risk management decisions. Cothern, et. al., using different high dose to low dose extrapolation models for evaluating the results from animal bioassays show that the predicted low dose results vary by a factor of 10^6; this they add "is like not knowing whether you have enough money to buy a cup of coffee or pay off the national debt."[6] Now the uncertainties may not be quite so drastic as their comments suggest, for some considerations of science may suggest at least the more extreme models may be ruled out,[7] but this example does indicate some of the possible extremes open to risk assessors. Thus, given such possibilities, if risk assessors say that as a result of their studies, risks are very low, e.g., a 10^{-6} increase in the lifetime risk of contracting cancer, then risk managers have no risks to regulate. On the other hand, if because of the great uncertainty, risk assessors are able to say that the risks are very high, e.g., greater than 10^{-3} increased lifetime risk of contracting cancer, risk managers may have a risk to manage, but the projected risk may be so high it in effect greatly limits their options.

B. Because of the potential uncertainties involved, others have argued that risk assessors can smuggle in their policy preferences based

6 Cothern, Coniglio and Marcus, *Estimating Risks to Health*, 20 ENVTL. SCI. AND TECH. 111 (1986).

7 Several federal agencies in their risk assessment policies argue that both scientific considerations and protection of public health favor adoption of the multistage high dose to low dose extrapolation model.

on their own individual judgments. While I think that is a possibility, the argument of this paper does not rest on this point.

C. However, a more important point is that since there are so many scientific uncertainties in the assessment of risks, agency scientists have to rely on inference assumptions to guide their decisions in at least four different kinds of circumstances:[8]

> (i)... when the data are not available in a particular case; (ii) assumptions potentially testable but not yet tested; (iii) assumptions that probably cannot be tested because of experimental limitations; (iv) assumptions that cannot be tested because of ethical considerations.

These assumptions or inference guidelines may reflect not only the best current guess about the correct science in question, but also social or political policy considerations, and it seems proper that they do so. However, when the policy considerations are introduced this blurs the distinction between the presumably "scientific" and "policy neutral" risk assessment and the policy laden risk management.

D. Finally, and most important for the purposes of this paper, even if the uncertainties did not exist and even if many of these inference guidelines could be replaced by correct scientific theories, there remain some features of risk assessment that necessarily will beg social policy questions in certain circumstances; these are a major focus of this paper.

Epidemiology is one tool used to estimate health risks to human beings from toxic substances. It is often difficult to do good epidemiological studies because there are *practical evidence gathering* problems.[9] Even when there are not, however, theoretical difficulties

---

[8] OFFICE OF TECHNOLOGY ASSESSMENT, IDENTIFYING AND REGULATING CARCINOGENS, 25 (1987).

[9] *See* discussion of this point in Cranor, *Epidemiology and Procedural Safeguards for Workplace Health Protection in the Aftermath of the Benzene Case*, 5 INDUS. REL. LAW J. 372 (1983).

undermine the policy neutrality of epidemiology and hence the policy neutrality of this aspect of cancer risk assessment. In particular, for the relatively rare diseases typical of many cancers and for small sample sizes, the design as well as the interpretation and use of epidemiological studies of such diseases depends on the use of controversial moral or social policy assumptions. Thus, I argue that in order to interpret and use the science necessary to estimate risks to human beings from exposure to toxic substances, scientists or the agency personnel interpreting the studies must incorporate the equivalent of moral or social policy assumptions as well as appropriate statutory preferences. If scientists follow their usual conventions in doing statistical studies, they will beg the legal questions at issue. In order to avoid this, our legal institutions should be modified in certain ways to take needed policy concerns into account. This argument is developed by reference to an elaborate example concerning the use of epidemiology.

II

Epidemiological studies are classified into experimental[10], descriptive[11], and observational studies. Observational epidemiology, the focus of this discussion, "depends on data derived from observations of individuals or relatively small groups of people."[12] Such studies are then analyzed with "generally accepted statistical methods to determine if an association exists between a factor and a

[10] Experimental studies require the "deliberate application of withholding of a factor and observing the appearance of lack of appearance of any effect." Use of such studies for purposes of discovering whether exposure to toxic substances can cause health harms would not be appropriate for obvious moral reasons. OFFICE OF TECHNOLOGY ASSESSMENT, CANCER RISKS 137 (1981).

[11] Descriptive epidemiology examines "the distribution and extent of disease in populations according to basic characteristics — e.g., age, sex, race, etc." These might provide clues to the etiology of disease which could then be investigated more thoroughly in other studies. *Id.*

[12] *Id.*

disease and, if so, the strength of the association."[13] Of the two species of observational studies — cohort and case-control studies — I consider only cohort studies.[14] Cohort studies can be retrospective or prospective. In a prospective study, a sample population exposed to a potential disease-causing factor is followed forward in time. Its disease rate is then compared with the disease rate of a group not similarly exposed. In a retrospective study, the same method is employed, only using historical data. Each kind of study has its advantages and its problems.[15] A thesis I have argued in more detail elsewhere[16] is that a wise and conscientious epidemiologist (or a risk manager using such studies) with perfect evidence, but with constrained sample sizes for detecting relatively rare diseases (with a background risk $< 10^{-4}$), *cannot* help but face potentially controversial moral and social policy decisions in order to interpret an epidemiological study and to produce the risk numbers that are the outcome of such work.

In trying to determine whether a substance such as a benzene is a

---

13 *Id.*

14 In a case-control study:

> people diagnosed as having a disease (cases) are compared with persons who do not have the disease (controls). Clearly, fewer people are needed in a case-control than in a cohort study, for only those with the disease, not those exposed to a risk factor, are the objects of examination. In either case, an association between a risk factor and the disease means that those exposed will tend to develop the disease and those not exposed will tend not to develop it. Case control studies are essentially retrospective. The researcher takes a group that has contracted a disease, compares the characteristics of that group and its environment with a nondiseased group, and tries to isolate factors that might have caused the disease.

J. MAUSNER & A. BAHN, EPIDEMIOLOGY: AN INTRODUCTORY TEXT 312-13, 322-25 (1974).

15 *See* J. KELSEY, W. THOMPSON, & A. EVANS, METHODS IN OBSERVATIONAL EPIDEMIOLOGY 128-130 (1986) for discussion of some of these problems.

16 Cranor, *Some Moral Issues in Risk Assessment* forthcoming in ETHICS (1990).

human carcinogen, a scientist considers two hypotheses. The first (the null hypothesis) predicates that exposure to benzene is *not* associated with greater incidence of a certain disease (e.g., leukemia or aplastic anemia) than that found in a nonexposed population. The second (the alternative hypothesis) indicates that exposure to benzene *is* associated with a greater incidence of such diseases.[17]

Since an epidemiological survey relies on statistical samples, by chance alone a researcher risks inferential errors from studying a sample instead of the whole population in question. In particular, one runs the risk of false positives (the study shows that the null hypothesis should be rejected and the alternative hypothesis accepted when in fact the null hypothesis is true) designated as a type I error,[18] or false negatives (the study shows that the null hypothesis should be accepted when in fact the null hypothesis is false and the alternative hypothesis is true), called a type II error (summarized in the table on the next page).[19]

Statistical theory provides estimates of the odds of committing such errors by chance alone. The probability of committing a type I error is normally designated α, and the probability of committing a type II error is designated β.[20] Conventionally, α is set at .05 so that there is only a one in twenty chance of rejecting the null hypothesis when it is true.[21] The practice of setting $\alpha = .05$ I call the "95% rule", for researchers want to be 95% certain that when the null hypothesis correctly characterizes the world, the statistical test will show the null hypothesis

17 A. FEINSTEIN, CLINICAL BIOSTATISTICS 320-21 (1977).

18 *Id.*, at 321-22.

19 *Id.*, at 324-25. Table I is adapted from *id.*, at 325 (Table I).

20 *See generally, id.*, at 320-34.

21 Walter, *Determination of Significant Relevant Risks and Optimal Sampling Procedures in Prospective and Retrospective Comparative Studies of Various Sizes*, 105 AM. J. EPIDEMIOLOGY 387, 391 (Table 2) (1977). I do not discuss how the various statistical variables are derived for a particular study from the raw data but only wish to show the conceptual relationships among them.

is correct and is accepted.

TABLE I

	Null hypothesis is actually true, e.g., benzene is not positively associated with leukemia.	Null hypothesis is false, alternative hypothesis is true, e.g., benzene is associated with leukemia.
Null hypothesis is accepted.	No error	Type II error False negative
Null hypothesis is rejected (and alternative hypothesis is accepted).	Type I error False positive	No error

Conventional practice is less rigid concerning values for β, but it is typical to set β between .05 and .20, when α is .05. The "power" of a statistical test is $1 - \beta$. When β equals .20, the power of one's statistical test is 0.80. This means a scientist has an eighty percent chance of correctly rejecting the null hypothesis as false when it *is* false. Choice of appropriate α and β values is guided by one's purposes; often these are philosophic in nature.

The low value for α for most scientific studies probably reflects a philosophy about scientific progress and may constitute part of its justification.[22] By keeping the chances of false positives quite low,

[22] The low value for α may also be a mathematical artifice explained historically. As Giere puts it:

> The reason [for the practice of having a 95% confidence level to guard against false positives] has something to do with the purely historical

then when one obtains a positive result one can have considerable confidence that one's addition to scientific knowledge is not the result of random chance. In building the edifice of science, by keeping the odds of false positives low, one ensures that each brick of knowledge added to the structure is solid and well-cemented to existing bricks of knowledge. Were one to tolerate higher risks of false positives, take greater chances of new knowledge being mistaken by chance alone, the edifice would be much less secure. A secure edifice of science, however, is not the only important social value at stake.

One can think of α, β, and $1 - \beta$ as measures of the "risk of error" or "standards of proof". What chance of error is a researcher willing to take? Is a twenty percent ($\beta = .20$) chance of saying benzene does not cause cancer, when in fact it might, an acceptable risk? When workers or the public may be contracting cancer (unbeknownst to all) even though a study (with high epistemic probability) shows they are not, is a risk to their good health worth a twenty percent gamble?

Alternatively, we might think of α, β, and $1 - \beta$ as standards of proof. How much proof do we demand of researchers and for what purposes? Must potential carcinogens be condemned by mere majority of the evidence, say somewhat more than fifty percent of the evidence (e.g., $1 - \beta = .51+$)? These questions only precede more complex matters, for the standards of proof demanded of statistical studies have implications for the costs of doing them and for the risks that *can be detected.* The mathematics of epidemiological studies, together with small sample sizes and rare diseases for study force serious policy choices on researchers and regulators alike when these studies are used

---

fact that the first probability distribution that was studied extensively was the normal distribution.

GIERE, UNDERSTANDING SCIENTIFIC REASONING, 212-213 (1981).

Two standard deviations on either side of the mean of a normal distribution encompasses 95% of the entire distribution.

in regulatory contexts to estimate risks to people.

The trade-offs at stake depend upon two other variables: N, the total number of people studied in the exposed and unexposed samples, and δ, the relative risk one wants to detect.[23] At the outset of the study, one might design a study so that δ is some value considered an unreasonable risk to health for public policy purposes, say a relative risk of 2, 5, or 10.[24] The value chosen depends upon many factors, including the seriousness of the disease, its incidence in the general population, and how great a risk, if any, the exposed group justifiably should be expected to run.[25] If one wishes to detect a very small relative risk between two groups in a cohort study, e.g., a relative risk of 2 for a rare disease, large numbers of exposed and unexposed individuals must be studied. A large relative risk, such as a risk of 6, requires fewer individuals to obtain statistically significant results.

The relation between the relative risk and sample size raises a more general issue. α, β, δ and N are mathematically interrelated. If any three of them are known the fourth can be determined. Because the variables are interdependent, crucial trade-offs are forced by the logic of the statistical relations. Consider the hypothetical decision tree (summarized in Table II) which presents five related alternatives.[26] The cohort study assumes that the prevalence of disease L in the general population is

---

23 *See* Feinstein, *supra* note 17, at 320-324.

24 That is, the incidence of disease in a group exposed to a risk factor would be two, five, or ten times greater respectively than the incidence of disease in the general population.

25 To be more precise about this, however, δ need not be set in advance, for it depends upon the number of people to be studied and the prevalence of the underlying disease, in addition to the values of α and β.

26 The numbers in the first and second alternatives are from Appendix A, Cranor, *Epidemiology and Procedural Protections for Workplace Health in the Aftermath of the* Benzene *Case*, 5 INDUS. REL. L. J. 372 (1983).

The numbers in the last three alternatives are from Appendix A of this paper.

8/10,000. It seeks to detect a relative risk of 3 ($\delta = 3$), provided such risks exist.[27]

From Table II, it is not immediately evident which alternative is the most attractive. Alternatives (1) and (2) would be excluded for reasons of cost or impracticality because the samples required are simply too large to be manageable even though these are the most accurate studies. Alternatives (3) and (4) may put those exposed to toxic substances at considerable risk, and alternative (5) risks undermining the credibility of the research because it is inconsistent with scientific practice, since it violates the 95% rule. The logic of epidemiology together with small sample sizes and a low background disease rate impose difficult moral choices on "scientific" research.

Furthermore, some risks may be statistically impossible to detect. Suppose, for social, regulatory or legal reasons, that it is thought important to detect a relative risk of 3 among workers exposed to toxic substances, for a disease that occurs in eight people of every 10,000. If there were only 1,000 workers to study (with α at .05 and β at .20), a relative risk could not be detected below 10, even if it turned out the substance in question did cause a threefold increase in mortality among workers.[28]

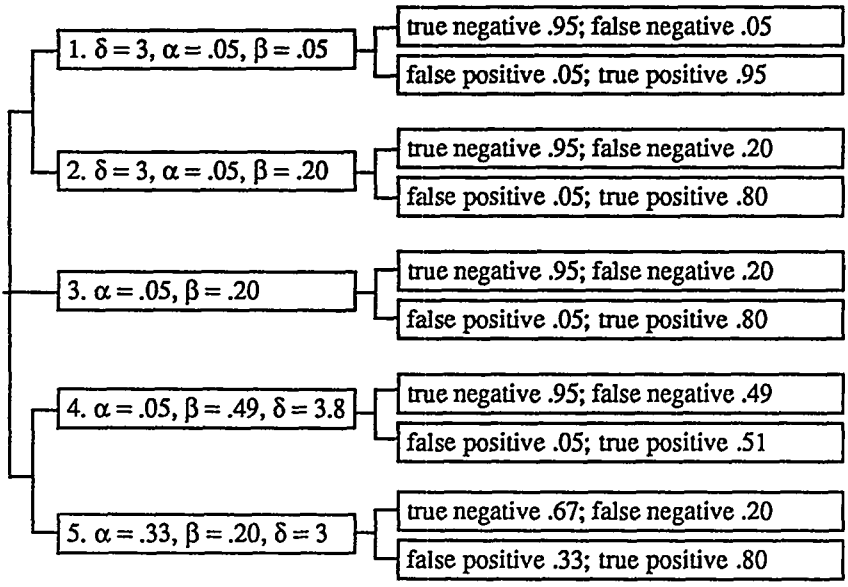
Alternative (3) suggests some interesting results for "negative" or "no effect" studies. Assume a study is run on 2,150 exposed workers with α at .05 and β at .20, when the prevalence of the underlying disease is 8/10,000. With these values, we only could be confident of detecting a relative risk of 6. But suppose no relative risk were detected, that is, the study was "negative" or showed "no effect" between the chemical C and the disease L. What could we infer? At most we would be justified in concluding that the relative risk was less than 6. It might

27 The alternatives are numbered in the left hand boxes, with H_0 being presented above H_1, in the right hand boxes, for each.

28 This figure is taken from Walter, *supra* note 21, at 391 (Table 2).

be 5.8 or 1, but given the constraints on the study, we could not conclude so statistically. In general, for "no effect" studies the most that can be inferred is that the relative risk to people in the exposed group is not as high as the relative risk tested for in the study.[29] Regulatory agencies regard such results as useful mainly for setting upper bounds on risks to people.[30]

TABLE II[31]



[29] This difficulty with negative human epidemiological studies also applies to animal bioassays which are essentially animal epidemiological studies. We should be similarly skeptical of "no effect" results there.

[30] *See supra* note 8, wherein the policies concerning use of epidemiological studies are summarized.

[31] For alternative 1, n/2 = 13,495.
For alternative 2, n/2 = 7,695.
For alternatives 3 to 5, n/2 = 2,150. With regard to (3) H_1, it can only be inferred that the relative risk is not as high as 6; with regard to (4) H_0 that odds are .49 that exposed subjects will remain exposed to harmful substances; and with regard to (5) H_1 that scientific credibility is undermined.

As striking as the preceding examples are, they only suggest the statistical problems a cohort study of a typical environmentally caused disease (e.g. benzene-induced leukemia), might pose, for they are based on the assumption that the prevalence of the hypothetical disease L in the general population is 8/10,000. If the prevalence of the disease were rarer by a factor of 10, which is typical of leukemia [32], then the decision tree would exhibit even more extreme results. These are summarized in Table III.

There, a cohort study is presented in which it is assumed that the prevalence of disease L in the general population is 8/100,000 and that a study seeks to detect a relative risk of 3 ($\delta = 3$). Five numbered alternatives[33] are shown (as in Table II), with H_0 and H_1 for each.

The upshot is that the rarer a disease, the greater the problems faced by epidemiologists, and the more acute are the tradeoffs imposed by the mathematics involved.[34]

---

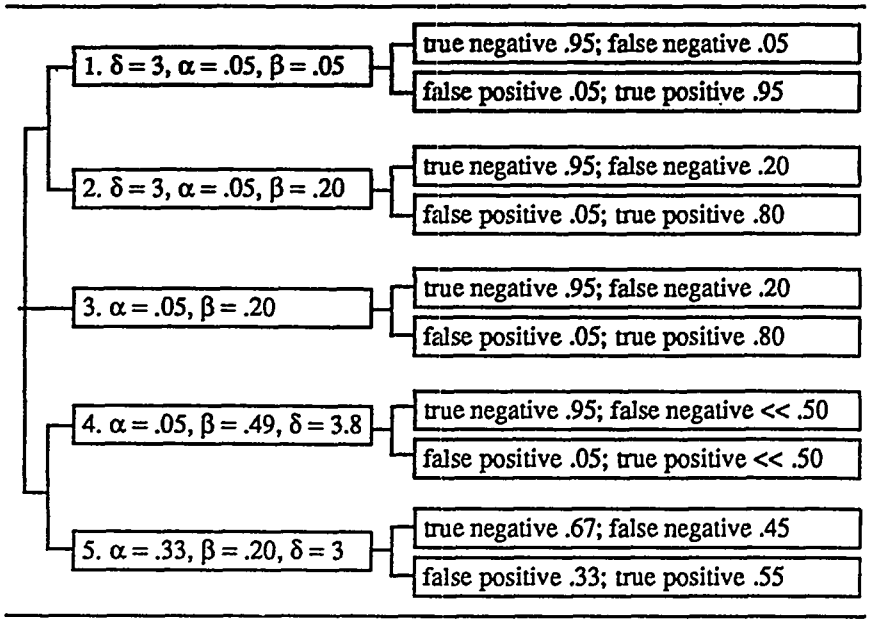
32 NATIONAL CANCER INST., DIV. OF CANCER CAUSES AND PREVENTION, DEMOGRAPHIC ANALYSIS SECTION, SURVEILLANCE, EPIDEMIOLOGY AND END RESULTS: INCIDENCE AND MORTALITY DATA 1973-1977, [MONOGRAPH NO. 57] 662-63 & Table 51 (1981).

33 The numbers for the first and second alternatives are taken from Appendix A in Cranor, *supra* note 26. The numbers for the fourth alternative appear, *id.*, at 392, note 111.

The numbers for the third and fifth alternatives appear in Appendix A of this paper.

34 The above problems are incident to a cohort study. A case-control study which looks only at diseased people and compares them with a control group requires fewer subjects, thus lowering the costs. The trade-offs and statistical difficulties imposed are exactly the same, however. The trade-offs involved between relative risk and type II errors mean that either study may conclude that people exposed to potentially toxic substances face no risk when in fact they do.

TABLE III[35]



III

The moral problems connected with epidemiological studies have other serious implications. How one *interprets* the fixed data of a study shows the value laden nature of the study. The fixed data, in a completed study consists of the background disease rate, sample size and revealed relative risk. For purposes of *interpreting* this information, epidemiologists (or risk managers using their studies)

[35] For alternative 1, n/2 = 135,191.
For alternative 2, n/2 = 77,087.
For alternatives 3 to 5, n/2 = 2,150. Also note with regard to (3) H_1 that 39 is the least significant relative risk which the study has .80 power to detect and with regard to H_0 that odds >> .5 that relative risk of 3.8 will not be detected when it exists. For alternative 5, note with regard to H_0 that there is high false negative rate and with regard to H_1 scientific credibility is undermined.

could vary the values of α and β. Consider one scenario as an example. Suppose the study of 2,150 exposed individuals revealed a relative risk of about 3, because there were 5 deaths compared with 1.72 (1 or 2) in the control group. Is this a positive result or not? The following table shows that one could interpret the study as a positive study for any of several pairwise choices of α and β values.

Any pairwise combinations of α and β in the left hand column will show that the study outcome is positive, for all would show a relative risk of about 3. Changing the variables slightly as indicated in the right column will produce a negative study. (We should also note that this study runs substantial risks of false negatives; these range from 49% down to 25%.)

Similarly, if the study revealed 6 deaths from exposure to a toxic substance, which is a relative risk of about 3.5 for an exposed group of 2,150, any pairwise combination of α and β values which enabled one to show a relative risk of at least 3.5 would produce positive results for the study. Changing the variables slightly would produce a negative study.

These examples show that epidemiologists, risk assessors and risk managers have considerable flexibility in *interpreting* the data of a study. How they interpret and *use* the data in certain regulatory and legal contexts will have important consequences for protecting human health.

TABLE IV

Positive Results*			Negative Results
α	β	$\delta^{\ddagger}$	
.10	.49	3.0	When α<.10 (with β constant) or β<.49 (with α constant)
.15	.40	3.0	When α<.15 (with β constant) or β<.40 (with α constant)
.20	.30	3.1	When α<.20 (with β constant) or β<.3 (with α constant)
Positive Results*			Negative Results
.25	.25	3.1	When α<.25 (with β constant) or β<.25 (with α constant)

*Least significant relative risk which the test has a power of .51 or higher to detect.

‡ Maximum observable relative risk.

Thus in some common circumstances (indicated above) in which we use or need to use the statistical tool of epidemiology, and in which scientific tradition would ordinarily require us to rely upon the 95% confidence rule, there is a tension between the use of this rule and other public policy and moral concerns we might have under environmental health laws. Roughly the tension is between a commitment on the one hand to traditional scientific caution in pursuit of the truth (represented in the 95% rule) and a commitment, on the other hand, to protecting people's health — or at least not taking chances with their health. However, the same examples show that in the circumstances described there is no necessity to the received scientific practice — it could be done differently. Whether statisticians and scientists should be

committed to the 95% rule in certain contexts is a normative, a policy, question that depends upon the purposes to which the results will be put. Thus, it raises substantial philosophical issues.

Second, the reporting of epidemiological data is not obviously a neutral and objective project. In the example just discussed, sample size and number of the deaths are fixed data in the study, but *whether a risk to human health is reported depends upon the choice of values for* α *and* β. How the fixed data gets used in subsequent regulatory or legal proceedings will also depend upon these variables and may have important consequences for our health. Whether scientists regard the study as presenting positive results or not, depends very much on the choice of values for α and β.

More importantly, the interpretation of the data is not value neutral, for, as we have seen above, the choice of values for α and β commits scientists (or those who use their results) implicitly, if not explicitly, to making judgments that are the equivalent of moral or social policy considerations. These equivalents of moral considerations must be relied upon in order to perform and interpret the studies in question. These choices are illustrated in the decision trees in Tables II and III. Thus, in a world of limited resources, scientists may be faced with a choice of spending large amounts of money in order to obtain *relatively precise results*, results which are scientifically respectable (α is low) and which have a small chance of false negatives (β is low) [for alternatives 1 and 2], or a choice of spending smaller amounts of money, but obtaining results that are *not scientifically respectable* (alternative 5) or results that have substantial odds of *producing false negatives by random chance* [alternatives 3 and 4]. These considerations enter into the design of the study.

Third, when there is an $\alpha - \beta$ asymmetry in testing large numbers of substances, there are also problems. As long as $\alpha < \beta$, and α is in

the neighborhood of .05, we are doing "better" science conventionally conceived, but as a matter of experimental design we may also be protecting possibly harmful chemicals better than human health. Suppose that we have 2400 substances to test. Assume, to be realistic, that 40% of those are carcinogens and 36% of them are not, with the remainder equivocal or inconclusive.[36] Now if epidemiologists set α at .05 and β at .20 (fairly typical values), assuming this is a large enough sample, we will have 192 false negatives and 43 false positives. With 192 false negatives, this means that our test will have falsely indicated that 192 substances were not carcinogenic when in fact they were. Thus, 192 substances which pose some risk of cancer to the populace (and how large a risk this is will depend upon both the prevalence of the disease, the relative risk associated with the substance, its potency and the number of people exposed) will not be detected. In addition, 43 false positives mean that 43 substances will be wrongly regulated (or possibly banned altogether), depending upon the statutory authority in question.[37] If substances are banned, the products into which they are incorporated will be more expensive to produce and market, or we will be deprived of their use and benefits altogether. If the substances are merely regulated, they will be more expensive to produce and market. Should it turn out that the percentage of carcinogens is less than 40%, the results would have to be similarly modified.[38]

---

36 OFFICE OF TECHNOLOGY ASSESSMENT, *supra* note 10, at 137.

37 If a substance falls under the Delaney Clause, *supra* note 2, and it causes cancer in one animal species, it will be banned according to a literal reading of the statute. (However, according to recent FDA interpretations of the Delaney Clause, this may not be true. If a substance causes a risk of cancer to animals that is so small that it is a de minimis risk, the FDA will not deem that it falls under the Delaney prohibition. *See* Correction of Listing of D & C Orange No. 17 for Use in Externally applied Drugs and Cosmetics, 52 F. R. 5081 (1987). More recently, the Circuit Court of Appeals for the District of Columbia has prohibited this FDA interpretation of the Delaney Clause. If it fell under other statutes, it might merely be regulated, or might escape regulation altogether, depending on the statute in question.

The points made above about human epidemiological studies are also applicable to the statistics of animal bioassays one of the bases of risk assessment. Talbot Page has shown[39] that in a bioassay with 50 controls and 50 experimental animals if the controls have 5 animals with tumors at a specific site (with 45 tumor free at that site) and 12 of the experimental (treated) animals have tumors at that site, using Fisher's exact test, the value for α for these results is .0542. Thus, the results are not statistically significant if one uses the 95% rule.

However, if one performs a Bayesian analysis on the same data and uses available prior information that historical controls have a 10% tumor rate, then one would greatly increase one's suspicion that the

38 How serious a problem is presented by the asymmetry between traditional α and β depends in part upon what moral theory you believe to be correct and the numbers of people exposed to the substance. It also is dependent upon the substances in question.

We might classify chemicals tested in an epidemiological study in two different ways: are the benefits provided by the substances comparable to the harms threatened or not? We might think of the benefits comparable to the harms, when, for example, one of the benefits promised is the saving of lives (comparable to the harm threatened by carcinogens which take lives) or the prevention of death. When the harms and benefits are comparable, then perhaps we should have a greater concern about the possibility of false positives, for we may lose substantial life saving benefits if we falsely condemn a substance when an epidemiological study has a false positive outcome.

Contrast this possibility with the results when we have a substance that does not promise life saving or death preventing benefits. For example, the Environmental Protection Agency recently issued an alert on Alar, a systemic chemical which is used on apples, peanuts, grapes and a number of other fruits. Its primary benefit in apples is to prevent them from falling off trees too early, to keep them firm and ripe-looking longer, and to delay the onset of rotting. The benefits are almost totally marketing and profitability benefits with no obvious beneficial health effects at all. The chief harm posed by Alar is that it may be one of the most potent carcinogens known, approaching the now banned EDB in potency. In an epidemiological study of Alar, using a small α value which ensures good science, the risks posed to our health may be great, because of the carcinogenic potency of the substance, but such risks might not be detected because of the chances of false negatives.

39 Page, *Problems with P-Values*, 2-3 (Manuscript in preparation).

substance was carcinogenic.[40]

The point: on traditional hypothesis testing with scientists using the 95% rule, this substance would be found to have "no toxic effect," because the results were not "statistically significant." Thus, use of the 95% rule may tend to give the substance a clean bill of health, although even statistically there appears to be substantial evidence of toxicity. This would be a mistake. Raising α, what counts as a statistically significant effect, would indicate there is evidence that the substance is toxic. Thus, use of the 95% rule might well wrongly lead to no regulation of this substance.[41]

Since the reporting, interpretation and use of epidemiological and animal bioassay data are not normatively neutral, and we could change conventional scientific practices, we should face the use of the 95% rule in these contexts as a normative question, as a legal and moral question. The choice of variables in an epidemiological study is a normative matter.

---

40 Thus, if our level of suspicion of the chemical's toxicity were a probability of .33, we would update our suspicion to a probability of .63, conditioned our observing the result of (5, 12) [tumors in controls, tumors in treated animals]; or if our initial level of suspicion were a probability of toxicity of .2, we would raise this probability to .46 on the basis of the evidence.

Id.

41 There is a generalization to the argument I have been offering. Any specialist (at least in academic disciplines) is concerned about the validity and defensibility of her inferences. The 95% rule is a common standard for good *statistical inferences.* By analogy with the arguments about epidemiology, to the extent that scientists are reluctant to conclude that suspect substances do not cause disease or death *because the inferences cannot be justified on the very best inference standards for their disciplines,* a debate whether to regulate or not may be begged in favor of non-regulation. This is most obvious in the case of statistical inferences in hypothesis testing, for the chances of false positives cannot be reduced without *increasing* the odds of false negatives. By analogy with the recommendations made above, scientists should similarly scrutinize other scientific inferences used in risk assessment to see whether regulatory outcomes are biased by scientific practices. Similarly, the use of strict scientific inferences in regulatory contexts should be addressed as moral or social policy questions.

Consider an analogy in the law. In criminal trials, avoiding wrongful damage to someone's reputation and well-being is so important that we impose quite demanding standards of evidence in order to establish guilt; we want to avoid wrongly inflicting harsh treatment and condemnation on the defendant (a legal false positive). We could save money and make proof of guilt easier if we thought it worth the human costs, but we do not. We have been quite self-conscious in debating the moral considerations that bear on the design and workings of the criminal law. The evidence problems in the interpretation and use of statistical human or animal studies that affect regulations which protect our health should receive similar treatment. In closing, I indicate a few of the considerations that bear on the use of such evidence the regulatory setting.

IV

The 95% rule should not be abandoned in all scientific contexts nor should it be abandoned in all regulatory contexts. However, since it is one standard of evidence, designed for certain purposes, scientists and agency personnel relying upon it should be discriminating in its use. In clinical trials of a drug in which the goal is to try to discover if a drug has therapeutic effects, it should be relied upon, for we should not conduct research endeavoring to add to fundamental knowledge about biochemical and therapeutic mechanisms which takes considerable chances of incurring false positives. Similarly, when one is conducting epidemiological research to establish knowledge as a foundation for further research, one might well want to retain the 95% rule.

In other contexts, however, it is likely there will be reasons for departing from the 95% rule:

— in preventive regulatory proceedings where the major concern is the forward-looking prevention of health harm, where there is little

fundamental research to be gained or upon which to build, and where the typical burden of proof imposed by the relevant statutes is typically lower than in the criminal or tort law, agencies should not adhere to it;[42]

— in screening substances to try to discover those that pose harms to health;

— and in the tort law where the typical standard of proof is not nearly as demanding as the 95% rule, perhaps courts should tolerate such departures.[43]

The point of these suggestions is that the law should carefully scrutinize any commitment to "scientific" standards of proof when using statistical studies, for such a commitment may beg the social policy questions at issue. Further, the appropriate standard of evidence to be applied in such cases should be dictated by the appropriate legal purposes and the applicable law (or wider social considerations).

V

In closing, I suggest some possibilities for regulatory law. First, I noted in section I that in doing cancer risk assessment, agencies must make decisions under conditions of substantial uncertainty. Second, some of the inference guidelines used to guide the decisions under

[42] For example, regulatory agencies may nominate substances to the National Toxicology Program for testing simply because they appear to pose health risks. The testing that is done under this program is aimed at regulatory purposes, and while some additional basic research may be obtained from the results of such tests, that is not the primary aim. (Although the research done under this program is typically experiments with animals, similar arguments would apply as we have seen above.) In fact, if the arguments of this paper are correct, perhaps regulatory agencies should not require adherence to the 95% rule for their purposes.

[43] Typically, the tort law requires for successful prosecution of a case that the plaintiff support his position by a "preponderance of the evidence." If this could be quantified, it appears to mean that more than 50% of the evidence favors the plaintiff — some place the estimate as high as 65%.

uncertainty are themselves chosen partly on policy grounds (section **I**). Third, the bulk of the paper (sections **II–IV**) has been devoted to indicating the *mathematical incompatibility* of simultaneously protecting against small chances of false positives and false negatives, the mathematical impossibility of both doing good science and good regulation when scientific evidence depends upon the statistics small samples. Thus, adherence to scientific standards of evidence in such cases may well beg the regulatory questions at issue before an agency.

Because of the observations in the above paragraph, there are reasons to use policy considerations to help guide agencies in addressing these problems. There are several sources of such policy considerations. A number of authors have suggested that the *statute* from under which an agency derives its authority ought to guide *risk assessment* choices and decisions under conditions of uncertainty and in the choice of inference guidelines. [44]

Hattis and Smith are typical. In proposing risk assessment guidelines they indicate that:[45]

> ...the guidelines must allow the analyst to select the particular form for expressing such uncertainty as may be relevant for choices under a specific statute.Under a risk/benefit balancing type of statute, the full probability density function for all sectors of the exposed population may be relevant to the decision-makers' choice, whereas only an "upper confidence limit" (at some defined probability level) for a select "sensitive subgroup" within the population may be relevant under a statute that requires the decision-maker to assure that the standard will "protect public health with an adequate margin of safety."

In addition, in regulatory contexts, the extent to which one might

---

[44] *See* Hattis & Smith, Jr., *What's Wrong With Risk Assessment?* in QUANTITATIVE RISK ASSESSMENT 95 (J. Humber & R. Almeder eds. 1986) and Latin, *Good Science, Bad Regulation, and Toxic Risk Assessment*, 5 YALE J. REG. 89 (1988), for two sets of authors who hold this view.

[45] Hattis & Smith, *supra*, at 95.

depart from the 95% rule in interpreting scientific evidence (and in distributing the costs of false negatives and false positives) should also be guided by the extent to which a statute explicitly mandates health protections. The more a statute requires health protections, the greater the departures from the 95% rule seems permitted (or even required).[46] The Delaney Clause of the Food, Drug, and Cosmetic Act (prohibiting direct food additives that cause cancer in humans or animals) and the hazardous substances section of the Clean Air Act might require greater departures than some other statutes. A much less demanding standard seems imposed by the health protections of the Federal Insecticide, Fungicide and Rodenticide Act which requires that the use of pesticides not generally cause "any unreasonable risk to man or the environment, taking into account the economic, social and environmental costs and benefits of the use of any pesticide."[47] This statutory language suggests that economic and other social costs may be balanced against potential health harms in deciding whether to permit a substance into commerce. Explicitly permitting wider social and economic costs to outweigh threats to health suggests that regulations under such a statute will be less protective than those under more risk averse statutes.

In addition, agencies can resort to statutory language to assign burdens of proof for evidence of causation. For this purpose it appears that regulatory agencies typically do not operate under as nearly demanding burdens of proof as those that exist in the criminal law or perhaps as demanding as those that exist in the tort law, for some statutes[48] seem to permit a lesser standard of proof and most others

---

46 Although it is not entirely clear why use of the 95% rule should be the default baseline from which departures should be argued for *regulatory* purposes. The 95% rule constitutes conventional scientific practice, but given the *regulatory* effect this rule can have, perhaps it should not be the default baseline.

47 7 U.S.C. § 136 (bb) (1982).

48 For example, the Occupational Safety and Health Act permits OSHA to regulate even when such regulations are on the "frontiers of scientific knowledge." However,

have to be administered under the "arbitrary and capricious" standard of the Administrative Procedure Act (APA).[49] The appropriate burden of proof, however, is imposed by the statute in question or implicitly by the APA — or court interpretations of the statute or the APA. Because of the lower burdens of proof under regulatory statues than in the tort law, for example, agencies should have less demanding standards of evidence for causation than the 95% rule used in peer reviewed journals.

Some authors have suggested further *neutral considerations* (that is, a consideration that is neutral between persons' views about the right outcome for regulation) for guiding risk assessments.[50] These include agencies' adopting policies consistent with (but not required by) the statutes in question,[51] avoiding the potential for catastrophic miscalculations,[52] and examining the effectiveness (on grounds of cost and other considerations) of their own programs so as to achieve the maximum protection per agency cost and effort.[53] In addition to these considerations, agencies should also be cautious about releasing sub-

---

the recent *Benzene* case may impose a somewhat more demanding burden on the agency. Industrial Union Dept., AFL-CIO v. American Petroleum Institute, 448 U.S. 607 (1980).

49 The "arbitrary and capricious standard" suggests that if an agency has not been arbitrary and capricious in interpreting evidence before it, it has wide discretion to act, possibly even though evidence does not measure up to the standards of evidence currently accepted in scientific disciplines.

50 Latin, *supra* note 44, is typical in this regard but is not the only author to propose such strategies.

51 Cranor, *Epidemiology And Procedural Protections For Workplace Health Protections In The Aftermath of the Benzene Case*, 5 INDUS. REL. L. J. 372, 396-399 (1983).

52 Latin, *supra* note 44, indicates this includes consideration of evidence of "widespread population exposures," "absence of a long historical record of exposures," and "evidence of unusual potency." (*Id.*, at 139-141). I would add other considerations to this: *evidence of substances that were particularly 'potent' in the environment because they have a long half-life, poor absorption rate in the soil, or are otherwise relatively inert in the environment but not in human or animal biological systems.*

53 Latin, *supra* note 44, at 138.

stances that have problematic properties in the environment: long degradation half-life, poor absorption rates in the soil, high solubility, high volitilization, relative inertness in the natural environment but not with mammals.

Finally, if statutory authority is sufficiently unclear and "neutral" considerations do not provide sufficiently definitive guidance, risk assessors can turn to broader moral and social considerations, to guide risk assessments such as the slogans that the regulatory law should be protective of the public health, should not risk especially sensitive subpopulations, etc. Such slogans are not comprehensive moral views at all, but merely hint at some appropriate moral considerations. More comprehensive moral views might include the following. A theory that places great weight on *protecting human health*, such as a *rights-based theory* might, would justify greater health protections, more cautious risk assessment procedures, and greater departures from the 95% rule in interpreting statistical studies than would a theory which places less weight on health protections, such as utilitarianism (the theoretical foundation of cost-benefits analysis) might. The attractiveness of a moral view that protects health as a matter of right is that the right to health care protections cannot be *trumped* or easily overridden by general social benefits (such as the costs for consumer goods, the benefits for the agricultural community or the total national product, etc.). Utilitarian (or cost-benefit) considerations typically have difficulty justifying health care protections by means of *rights* because they permit general social considerations to override specific individual protections. The point of this is that when statutory authority and "neutral" principles for guiding risk assessment choices are exhausted (if they are), agencies may also turn to more general moral and political considerations similar to these in order to guide their decisions.

A final point. Inescapably statutory guidance, neutral considerations

and broader moral considerations must be used to guide risk assessment procedures which threatens the "scientific neutrality" of risk assessment. This suggestion may invite criticism from the scientific and the regulatory (or regulated) community. Nonetheless, it seems that such considerations must be appealed to for the reasons given in this paper. The only issue is how it should be done. If such considerations are *surfaced, consciously debated,* and *publicly affirmed and adopted* after debate, in a democracy this is an appropriate (even required) role for public participation in risk assessment and the regulatory decisions that affect peoples' lives. *Public participation* in adopting policy considerations that guide risk assessment decisions under uncertainty and that guide the distribution of regulatory false positives and false negatives seems required if such policy considerations are going to have the important role they must have in risk assessment. If risk assessment at present is inescapably permeated with policy considerations, as I have argued it is, then the policies that guide the process should be the outcome of substantial public participation as is appropriate in a democratic form of government.

APPENDIX A
Relative Risk as a Function of Alpha and Beta Values

I

Relative Risk [Rel.Risk] When Disease Rate [Dis. Rate] is 8/10,000

	Alpha	Beta	Dis. Rate	Rel. Risk	Sample[54]
1	0.05	0.05	8×10^{-4}	8.9	2150
2		0.10		7.5	2150
3		0.15		6.7	2150
4		0.20		6.0	2150
5		0.25		5.5	2150
6		0.03		5.1	2150
7		0.35		4.7	2150

54 This is one-half of the total population.

	Alpha	Beta	Dis. Rate	Rel. Risk	Sample
8	0.05	0.40	8×10^{-4}	4.3	2150
9		0.45		4.0	2150
10		0.49		3.8	2150
11	0.10	0.05		7.5	2150
12		0.10		6.2	2150
13		0.15		5.5	2150
14		0.20		4.9	2150
15		0.25		4.5	2150
16		0.30		4.1	2150
17		0.35		3.8	2150
18		0.40		3.5	2150
19		0.45		3.2	2150
20		0.49		3.0	2150
21	0.15	0.05		6.7	2150
22		0.10		5.5	2150
23		0.15		4.8	2150
24		0.20		4.3	2150
25		0.25		3.9	2150
26		0.30		3.6	2150
27		0.35		3.2	2150
28		0.40		3.0	2150
29		0.45		2.7	2150
30	0.20	0.05		6.0	2150
31		0.10		4.9	2150
32		0.15		4.3	2150
33		0.20		3.8	2150
34		0.25		3.4	2150
35		0.30		3.1	2150
36		0.35		2.8	2150
37		0.40		2.6	2150
38		0.45		2.4	2150
39	0.25	0.05		5.5	2150
40		0.10		4.5	2150
41		0.15		3.9	2150
42		0.20		3.4	2150

43	0.25	0.25	8×10^{-4}	3.1	2150
44		0.30		2.8	2150
45		0.35		2.5	2150
46		0.40		2.3	2150
47		0.45		2.1	2150
48	0.30	0.05		5.1	2150
49		0.10		4.1	2150
50		0.15		3.5	2150
51		0.20		3.1	2150
52		0.25		2.8	2150
53		0.30		2.5	2150
54		0.35		2.2	2150
55		0.40		2.0	2150
56		0.45		1.8	2150
57	0.33	0.05		4.9	2150
58		0.10		3.9	2150
59		0.15		3.4	2150
60		0.20		2.9	2150
61		0.25		2.6	2150
62		0.30		2.4	2150
63		0.35		2.1	2150
64		0.40		1.9	2150
65		0.45		1.7	2150

II

Relative Risk [Rel. Rate] When Disease Rate is 8/100,000

	Alpha	Beta	[unexposed]	Rel. Risk	Sample
1	0.05	0.05	8×10^{-5}	65.9	2150
2		0.10		52.6	2150
3		0.15		44.8	2150
4		0.20		38.8	2150
5		0.25		34.1	2150
6		0.30		30.4	2150
7		0.35		26.9	2150
8		0.40		23.8	2150
9		0.45		21.2	2150

	Alpha	Beta	[unexposed]	Rel. Risk	Sample
10	0.10	0.05	8×10^{-5}	52.6	2150
11		0.10		40.9	2150
12		0.15		34.1	2150
13		0.20		28.9	2150
14		0.25		24.9	2150
15		0.30		21.8	2150
16		0.35		18.9	2150
17		0.40		16.4	2150
18		0.45		14.3	2150
19	0.15	0.05		44.8	2150
20		0.10		34.1	2150
21		0.15		28.0	2150
22		0.20		23.4	2150
23		0.25		19.8	2150
24		0.30		17.1	2150
25		0.35		14.5	2150
26		0.40		12.4	2150
27		0.45		10.6	2150
28	0.20	0.05		38.8	2150
29		0.10		28.9	2150
30		0.15		23.4	2150
31		0.20		19.2	2150
32		0.25		16.0	2150
33		0.30		13.6	2150
34		0.35		11.4	2150
35	0.20	0.40		9.5	2150
36	0.33	0.05		28.2	2150
37		0.10		20.0	2150
38		0.15		15.5	2150
39		0.20		12.2	2150
40		0.25		9.8	2150
41		0.30		8.0	2150
42		0.35		6.4	2150
43		0.40		5.1	2150
44		0.45		4.1	2150